CIRRELT

Centre interuniversitaire de recherche
sur les réseaux d'entreprise, la logistique et le transport

Interuniversity Research Centre
on Enterprise Networks, Logistics and Transportation

# The Ambulance Relocation and Dispatching Problem

**Valérie Bélanger**
**Ettore Lanzarone**
**Angel Ruiz**
**Patrick Soriano**

**November 2015**

**CIRRELT-2015-59**

Document de travail également publié par la Faculté des sciences de l'administration de l'Université Laval, sous le numéro FSA-2015-014.

**Bureaux de Montréal :**
Université de Montréal
Pavillon André-Aisenstadt
C.P. 6128, succursale Centre-ville
Montréal (Québec)
Canada H3C 3J7
Téléphone : 514 343-7575
Télécopie : 514 343-7121

**Bureaux de Québec :**
Université Laval
Pavillon Palasis-Prince
2325, de la Terrasse, bureau 2642
Québec (Québec)
Canada G1V 0A6
Téléphone : 418 656-2073
Télécopie : 418 656-2624

www.cirrelt.ca

Université Laval · McGill · Concordia · ÉTS · UQÀM · HEC Montréal · Polytechnique Montréal · Université de Montréal

# The Ambulance Relocation and Dispatching Problem

**Valérie Bélanger[1,2,*], Ettore Lanzarone[3], Angel Ruiz[1,4], Patrick Soriano[1,2]**

[1] Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation (CIRRELT)

[2] Department of Management Sciences, HEC Montréal, 3000 chemin de la Côte-Sainte-Catherine, Montréal, Canada H3T 2A7

[3] Institute of Applied Mathematics and Information Technology (IMATI), National Research Council of Italy (CNR), Via Bassini 15, 20133, Milan, Italy

[4] Department of Operations and Decision Systems, 2325 de la Terrasse, Université Laval, Québec, Canada G1V 0A6

**Abstract.** Emergency medical services (EMS) generally deal with two real-time decisions: ambulance dispatching and relocation. Dispatching consists in selecting which ambulance to send to an emergency call, while relocation consists in determining how to modify the location of available ambulances throughout the day in response to changes in the state of the system. Although they have been mostly considered separately both in the literature and among practitioners, dispatching and relocation decisions are closely related. Considering them simultaneously could help improve the service level with lower relocation efforts. In this study, we address the Ambulance Relocation and Dispatching Problem (ARDP). The ARDP determines the location of each available ambulance as well as a dispatching policy, aiming at minimizing the expected response time along with relocation efforts. The ARDP is formulated as a linear programming model. To solve real-life instances, a matheuristic decomposition approach is developed exploiting the division of the territory into subregions. Results show the benefits of considering dispatching and relocation decisions simultaneously, and their impact on each other. Moreover, on a set of instances representative of a real application, the proposed approach has proven to be an effective tool to provide good quality solutions.

**Keywords**. Emergency medical services, relocation, dispatching, matheuristic decomposition.

**Acknowledgements.** This research was funded by the Fond de recherche du Québec - Nature et les technologies (FQRNT) through the Team research project program [grant PR-122269] and by the Natural Sciences and Engineering Research Council of Canada (NSERC) through the Discovery grants program [grants OPG 0293307 and OPG 0177174] and the Alexander Graham Bell Canada Graduate Scholarships-Doctoral Program awarded by the first author. We want to thank both organizations for their support.

Results and views expressed in this publication are the sole responsibility of the authors and do not necessarily reflect those of CIRRELT.

Les résultats et opinions contenus dans cette publication ne reflètent pas nécessairement la position du CIRRELT et n'engagent pas sa responsabilité.

---

* Corresponding author: Valerie.Belanger@cirrelt.ca

Dépôt légal – Bibliothèque et Archives nationales du Québec
Bibliothèque et Archives Canada, 2015

© Bélanger, Lanzarone, Ruiz, Soriano and CIRRELT, 2015

## 1. Introduction

Emergency medical services (EMS) are responsible of providing medical assistance to any person requiring it at the scene of an emergency as well as transportation to a health facility when necessary. The ultimate goal of EMS is to save lives and minimize the effect of health-related incidents. Once a call is received, the ambulance to assign to the call need to be selected quickly in order to reach the emergency scene and to provide help to the patient in a timely manner. The common rule that always sends the nearest idle ambulance is highly considered both among EMS organizations and the scientific community (Bandara et al., 2014; Jagtenberg et al., 2015). However, under given circumstances, it may be more effective to choose another ambulance that the nearest one in order to reduce the impact on the system and its ability to serve future demands, or to prevent workload imbalance (Toro-Diaz et al., 2013).

On the other hand, the location of ambulances also influences the time needed to reach an emergency scene. Location decisions are therefore an important, yet challenging, part of EMS management. Over a workday, ambulance locations can either be fixed, meaning that every ambulance will be sent back to its home base after completing a mission, or modified according to the system's state. Indeed, the arrival of emergency calls being highly uncertain, it may happen that at a given time, the ambulances available to answer these calls are no longer able to properly serve all regions, even if their location was carefully planned initially. To alleviate the stochastic and dynamic nature of emergency calls, relying on more flexible management strategies, such as ambulance relocation, can help improve the service level (Bélanger et al., 2014). Despite the potential service improvement, relocation strategies also generate ambulance movements that contribute to an increase in undesirable consequences from both economical and human resources management standpoints. These undesirable consequences, or relocation costs, should be carefully taken into account in the decision-making process.

Although the relationship between location and dispatching decisions is rather intuitive, both types of decisions are generally considered independently. Most of the studies dealing with relocation assume that the nearest ambulance is always dispatched to an emergency call. On the other hand, studies dealing with dispatching decisions generally consider that ambulance locations are known and that each ambulance return to its home base after completing a mission. To the best of our knowledge, Toro-Diaz et al. (2013) were among the first to propose a joint dispatching and location strategy. They proposed a model that determines simultaneously ambulance locations and an ordered list of ambulances to dispatch to each demand zone. To deal with the stochastic nature of emergency demands, they embedded a queuing model, the hypercube model (Larson, 1974), within a mathematical programming one, leading to a complex and challenging problem to solve when applied in a real-life context. To overcome this difficulty, Toro-Diaz et al. (2015) developed a tabu search metaheuristic to solve larger size instances. However, in both cases, the problem solved by Toro-Diaz et al. (2013, 2015) is considered at the strategic/tactical level: no relocation movements nor costs are taken into account.

In this paper, we define the Ambulance Relocation and Dispatching Problem (ARDP), which determines the location of each available ambulance as well as an ordered list of available ambulances for each demand zone, i.e. the pre-assignment list, or alternately, the priority list, that will guide dispatching decisions. The objective of this model is to minimize the expected response time along with relocation efforts. The ARDP thus allows to explicitly account for relocation inconvenience. The above mentioned problem is formulated as a mathematical model. Unlike Toro-Diaz et al. (2013), the ARDP considers the uncertainty related to ambulance availability through the calculation of the expected response time and the expected workload. The underlying mathematical model consequently remains linear: small and medium-size instances can be solved by the means of a commercial solver. Nonetheless, a solution approach is needed to solve real-life instances more efficiently. To do so, we developed a matheuristic decomposition approach exploiting the division of the territory into subregions. Each step of the methodology consists of a linear mathematical model extracted from the global model. Each subproblem is solved using a commercial solver, in our case, by the means of CPLEX 12.5.

The rest of the paper is organized as follows: Section 2 briefly discusses the literature related to the problem under study. In Section 3, the ARDP is presented, defined and formulated. The matheuristic decomposition approach developed to solve the ARDP is described in Section 4. Computational experiments are reported and analyzed in Section 5. Finally, a discussion on potential research avenues is reported in Section 6.

## 2. EMS related literature

A lot of effort has been devoted to many aspects of EMS management, e.g. demand forecasting, crew scheduling, performance measure, ambulance location (Ingolfsson, 2013). The problem we are dealing with in this paper is related to two main decision problems, namely the dispatching and the relocation. In this section, we briefly describe what we deem to be the most relevant works related to these two specific topics. We refer the reader to the surveys presented by Brotcorne et al. (2003), Goldberg (2004), Başar et al. (2012) and Bélanger et al. (2012) for a more detailed description of EMS related problems, and in particular, ambulance location ones. We also refer the reader to Bélanger et al. (2015) for a review of recent advances in EMS management, where a special interest has been dedicated to relocation problems.

### *2.1. Dispatching rules*

The nearest idle ambulance policy, which always sends the closest ambulance to serve an emergency call, is widely considered, both in the scientific community and among EMS organizations. Indeed, from a practical standpoint, this strategy is very easy to implement, but more importantly, it guarantees a rapid intervention for all emergency calls. The nearest idle policy is generally adopted for all cases, from the most urgent to the less prioritized ones. However, Carter et al. (1972) demonstrated that this dispatching policy is not always optimal to minimize the average response time, defined as the time elapsed from the arrival of a call to the arrival of a paramedical team at the emergency scene. Schmid (2012) and Bandara et al. (2014) also showed that, in the case of less prioritized calls, the nearest idle strategy may not be the best to adopt. Although it intends to minimize the response time to reach a call, this myopic strategy does not take into account vehicles' non-availability and its impact on the capacity of the system to serve future demands. To overcome this issue, other dispatching policies can be envisioned. Those policies should better consider the impact of sending a specific ambulance to serve a demand on future system performances while ensuring that each call can still be reached within a given time frame.

Following this idea, Gendreau et al. (2001) proposed to dispatch the idle ambulance that can reach the emergency scene within a prescribed time frame so that relocation costs are minimized. To select the right ambulance to dispatch, they computed the relocation plan corresponding to each possible dispatching decision using the time available between the arrival of two consecutive emergency calls. Once the call is received, the best ambulance to dispatch is chosen given the relocation plan associated with each eventual dispatching decision. Andersson and Värbrand (2007) suggested to dispatch the idle ambulance that can reach the emergency scene within a prescribed time frame and that will incur the smallest preparedness degradation, defined as the capacity of the system to serve future demands. Schmid (2012) proposed to adopt a dispatching policy based on an approximate dynamic programming model to minimize the average response time. McLay and Mayorga (2013) and Bandara et al. (2012) used Markov decision process approches to obtain optimal dispatching policies that seek to maximize, respectively, the coverage level and the patient survival. A dispatching policy that integrates the severity of a call to increase survival probability of patients is also proposed in Bandara et al. (2014). Their heuristic suggested to dispatch the nearest idle ambulance to high priority calls and the less busy ambulance for low priority calls. Finally, as discussed in introduction, Toro-Diaz et al. (2013, 2015) proposed a joint location and dispatching problem. In this study, dispatching decisions are taken according to fixed preference lists, in a similar way as what we propose in this paper. Their results showed the potential of dealing with both decisions simultaneously, especially when other performance indicators are considered, e.g. workload balance.

### *2.2. Relocation strategies*

Ambulance relocation has received much more attention than dispatching, but still a few compared to its static location counterpart (Brotcorne et al., 2003; Bélanger et al., 2015). Ambulance relocation is a strategy that modifies ambulance locations, or alternately standby sites, through a day to better adapt to the system's evolution. *Multi-period* relocations are performed at given and fixed times over a day. A workday is then divided into time periods, and location plans are defined for each time period. *Dynamic* relocations occur at specific, yet unknown, times at which the system state justifies it. In this case, two strategies can be adopted. First, relocation decisions can be determined for each possible state. Doing so, a set of look-up tables, defining the location of ambulances for each possible state, are known a priori and applied at the right time. Second, relocation decisions can be selected in real-time, based on the current system state. Main multi-period relocation approaches are reported in section 2.2.1 whereas dynamic relocation approaches, from both perspectives, are discussed in section 2.2.2.

*2.2.1. Multi-period relocations*

To the best of our knowledge, Repede and Bernardo (1994) formulated the first multi-period ambulance location model. Their probabilistic model sought to maximize the expected coverage while considering variations in both the demand pattern and the number of available ambulances. Relocation costs incurred between periods are however not taken into account in that model. Start-up and relocation costs have later been included in the model by van den Berg and Aardal (2015). Rajagopalan et al. (2008) presented another probabilistic multi-period model, which consists in finding the minimum number of ambulances required to guarantee that each demand zone is covered with a given level of reliability. Saydam et al. (2013) extended the latter model to include the minimization of the number of relocated ambulances as a second objective. Başar et al. (2011) considered the problem of determining the location and time at which ambulance stations (as opposed to standby sites) need to be open over a multi-period planning horizon. The proposed model limits the number of stations used at each time period while ensuring that all the population can be covered by two distinct stations within a given time frame. Finally, Schmid and Doerner (2010) presented a multi-period model that considers travel time variations between periods, due, for instance to road traffic congestion. A penalty term is also included in the objective function to limit the number of ambulances that need to be relocated between periods.

*2.2.2. Dynamic relocations*

The first model that explicitly addressed dynamic ambulance relocation was proposed by Gendreau et al. (2001). In this case, the authors suggested to solve the relocation problem each time an ambulance is sent to serve an emergency call. More precisely, their model aimed to determine ambulance locations that maximize the population covered by at least two ambulances within a given time frame. It also simultaneously sought to minimize relocation costs, which are directly related to each ambulance's relocation history. Gendreau et al. (2006) later proposed another dynamic relocation model, but specifically dedicated to the location of physician cars, instead of ambulances. This model determines the best possible physician car locations such that the expected coverage for each possible system state, defined as the number of available cars, is maximized. The model is solved a priori providing the decision maker with a set of look-up tables, one for each state. Andersson and Värbrand (2007) rather proposed to regularly check the preparedness level and to trigger ambulance relocation each time it drops below a given threshold. The proposed model is then solved to regain a minimal preparedness level for each demand zone. Following the idea of Gendreau et al. (2006), Nair and Miller-Hooks (2009) formulated a multi-objective location-relocation model that aims at maximizing the double coverage as well as to minimize location-relocation costs. The latter model also provides a set of look-up tables, one for each possible state. Naoum-Sawaya and Elhedhli (2013) developed a two-stage program to handle a priori relocation decisions: first stage decisions deal with the location of ambulances, and second stage decisions, with the assignment of emergency demands to ambulances. The model thus sought to minimize the number of relocated ambulances and the number of demands that cannot be reached within a given time frame. Mason (2013) proposed a dynamic ambulance relocation problem, that he embedded into an EMS management software. The problem considered in this case shares a lot of similarities with the one of Gendreau et al. (2001). Finally, Sudtachat et al. (2016) suggested a nested-look-up table policy where a limited number number of relocations can occur at the same time. The objective of this policy is to maximize the expected coverage.

In all previous cases, relocation decisions involve all available ambulances. However, in some contexts, it is not possible to relocate ambulances, for legal or computational matters. Consequently, some authors proposed to dynamically select the location of a newly idle ambulance, modifying the system one ambulance at a time, instead of considering the relocation of all available ambulances at the same time. Following this idea, Maxwell et al. (2009) and Schmid (2012) relied on approximate dynamic programming to formulate ambulance relocation problems that are limited to the newly idle ambulance. Both problems seek, in their own ways, the best possible service level. Jagtenberg et al. (2015) also defined a policy that sends the newly idle ambulance to the location or standby site that results in the largest marginal coverage.

This brief literature overview shows that dispatching and location problems are generally addressed independently. As discussed in the introduction, except from the works of Toro-Diaz et al. (2013, 2015), it seems that joint dispatching and location strategies have not been studied extensively. Moreover, it is important to recall that these works have been considered in a static location context where no relocation movements nor costs are taken into account. Nevertheless, those aspects can be important to consider in real-life context.

This therefore raises the need for the definition and the modelling of a decision problem that simultaneously addresses dispatching and relocation decisions, in a framework that can be used to solve real-life problems. The analysis of such a problem will also allow us to highlight the relationship between dispatching and relocation decisions, as well as the impact of considering relocation costs on decision-making. In the next section, we will present the definition of the problem under study along with its main characteristics and formulation.

## 3. The ambulance relocation and dispatching problem

The ambulance relocation and dispatching problem (ARDP) seeks to determine the location of available ambulances as well as the best dispatching policy defined as a set of pre-assignment lists, one for each demand zone. Each pre-assignment list consists of a list of ambulances ordered according to their priority of dispatch to calls arising in a given demand zone. Accordingly, if an emergency call arises in a zone, the first idle ambulance on the list is dispatched to the call. If there is no idle ambulance on the list, the nearest one will be dispatched to the call. If there is no idle ambulance at all, the call will be queued or redirected to another emergency organization. In addition, the ARDP includes relocation costs that are incurred when ambulances are moved from their current standby sites to another. In a real-life application context, the ARDP can be solved at predetermined time, or when the system state justifies it, for instance, when the expected performance drops below a given threshold as suggested in Andersson and Värbrand (2007).

Three main characteristics, considered all together, differentiate the ARDP from other problems. First of all, the ARDP suggests a joint strategy to consider both location and dispatching decisions with the objective of maintaining an adequate service level with lower relocation efforts. The elaboration of pre-assignment lists will help guide dispatching decisions. It will allow to take into account future dispatching decisions when looking for the best location plan. The ARDP also considers the capacity of ambulances or, alternatively, their maximal workload. Considering pre-assignment lists will ensure a proper computation of the expected workload assigned to each ambulance thus allowing the formulation of workload constraints. Finally, the expected response time, rather than the traditional coverage measure, is used to assess system performances. Both the expected response time and the expected workload are defined in a similar way as the widely used expected coverage (Daskin, 1983). However, although the expected coverage is not as precise as queuing theory-based model to estimate system performances under uncertainty, it offers the opportunity of constructing a model that is easier to understand by practitioners and easier to solve when facing larger size problems. Both conditions are often required for our solutions to be adopted in real-life contexts. The ARDP will also include the relocation time as a measure to assess relocation costs.

The dynamic ambulance relocation and pre-assignment problem can be defined on a graph $G = (V, E)$ where $V = I \cup J$, $I = (v_1, ..., v_n)$ and $J = (v_{n+1}, ..., v_{n+m})$ are two vertex sets representing, respectively, demand zones and potential standby sites, and $E = \{(v_i, v_j) : v_i, v_j \in V\}$ is the edge set. A demand zone is defined as a population area to which correspond a centroid and a population density, or alternately, a demand. A potential standby site is defined as a physical location where one or many ambulances can be sited while waiting to answer to emergency calls. A travel time $t_{ij}$ is associated to each edge $(v_i, v_j) \in E$ and a population density or a demand $d_i$ to each vertex $v_i \in I$. The maximal number of ambulances $p_j$ that can be located at standby site $j$ at the same time also corresponds to each vertex $v_j \in J$. In addition to these parameters, we define $K$, the set of available ambulances, and $Z$, the set of positions on pre-assignment lists, where $|Z|$ is in fact the number of ambulances included in pre-assignment lists. A parameter $\lambda_j^k$ that is set to 1 if ambulance $k$ is located to standby site $j$ prior to the relocation, 0 otherwise, and a parameter $\theta^k$ that is set to 1 if the relocation of ambulance $k$ is included in the computation of relocation costs, 0 otherwise, are associated to each ambulance $k \in K$. We also use $q$, the busy fraction of ambulances defined as in Daskin (1983), and $W$ the capacity or the maximal workload that can be assigned to an ambulance in terms of the number of interventions for a given time period. Both the busy fraction and the capacity are assumed to be the same for all ambulances. Finally, $\xi_s$ and $\xi_r$ correspond to the weights given to service level and relocation costs in the objective function, with $\xi_s + \xi_r = 1$.

### *3.1. Expected response time definition*

As we discussed previously, two main criteria are included in the ARDP objective function, both of which predict the service to the population and relocation efforts required to achieve such a service level. The expected response time is used to assess how good will perform the system, and the relocation time is defined as a relocation cost measure. Before going on with the ARDP formulation, it is worth discussing briefly the

expected response time measure. To compute the expected response time for a given demand zone, firstly, it is needed to know the location of ambulances and which ambulance will be sent to intervene in a given demand zone, i.e. the pre-assignment lists. The "real" expected response time will also include the probability that there is no idle ambulance on the list, in which case the nearest idle ambulance is sent to the call, and the probability that no ambulance is idle at all, in which case the call is queued or redirected to another emergency service. In the latter case, the response time is assumed to be known and fixed to $T$. Given $K_Z^i$, the set of ambulances on the pre-assignment list of a demand zone located at $v_i \in I$ ordered according to the pre-assignment list, where $z_k$ is the position of ambulance $k$ on the list, and $K_A^i$, the set of ambulances that are not on the pre-assignment list ordered according to an increasing response time criteria $t_{ki}$, where $a_k$ is the position of ambulance $k$ on this ordered list, the expected response time for calls arising in demand zone $v_i \in I$, noted $ERT_i$ can be estimated by:

$$ERT_i = \sum_{k \in K_z^i} (1-q) q^{z_k - 1} t_{ki} + \sum_{k \in K_a^i} (1-q) q^{|Z| + a_k - 1} t_{ki} + q^{|K|} T. \tag{1}$$

Although the ARDP only includes the first term in its objective function, mainly for computational reasons, the "real" expected response time should be computed and considered during the model analysis. This is particularly important when comparing scenarios with varying numbers of ambulances or list sizes.

| | |
|---|---|
| *Sets* | |
| $I$ | Demand zones |
| $J$ | Potential standby sites |
| $K$ | Ambulances |
| $Z$ | Position on pre-assignment lists |
| *Parameters* | |
| $t_{ij}$ | Travel time |
| $d_i$ | Population density or demand |
| $p_j$ | Maximal number of ambulances |
| $q$ | Busy fraction |
| $W$ | Maximal workload of ambulances |
| $\lambda_j^k$ | Ambulance $k$'s location prior to relocation |
| $\theta^k$ | Ambulance $k$'s impact on relocation costs |
| $\xi_s$ | Weight given to service level in the objective function |
| $\xi_r$ | Weight given to relocation costs in the objective function |
| $T$ | Response time when referred to another service |
| *Variables* | |
| $x_{j_1 j_2}^k$ | Ambulance relocation variables |
| $w_i^{zk}$ | Assignment variables |
| $y_{ij}^{zk}$ | Linking variables |

Table 1: Sets, parameters and variables

### *3.2. Formulation*

The ARDP deals simultaneously with two types of decisions: location decision assigns a standby site to each available ambulance, and pre-assignment decision defines an ordered list of ambulances for each demand zone. To address those decisions, three sets of variables have been defined: $x_{j_1 j_2}^k$ is a binary variable that is set to 1 if ambulance $k$ moves from standby site $v_{j_1} \in J$ to standby site $v_{j_2} \in J$ after relocation, 0 otherwise (if $j_1 = j_2$, no relocation is required); $w_i^{zk}$, a binary variable that is set to 1 if ambulance $k$ is at the $z$-th position on the pre-assignment list defined for the demand zone located in $v_i \in I$, 0 otherwise; and $y_{ij}^{zk}$, a binary variable that is set to 1 if ambulance $k$, located at $v_j \in J$ after relocation, is at $z$-the position on the pre-assignment list defined for the demand zone located in $v_i \in I$, 0 otherwise. Sets, parameters and variables are summarized in Table 1.

Given these sets, parameters and variables, the ARDP can be formulated as follows:

$$\min \xi_s \sum_{i \in I} \sum_{z \in Z} \sum_{k \in K} \sum_{j \in J} (1-q) q^{z-1} d_i t_{ji} y_{ij}^{zk} + \xi_r \sum_{k \in K} \sum_{j_1 \in J} \sum_{j_2 \in J} \theta^k t_{j_1 j_2} x_{j_1 j_2}^k \tag{2}$$

subject to:

$$\sum_{j_1 \in J} \sum_{j_2 \in J} \lambda_{kj_1} x_{j_1 j_2}^k = 1, \ \forall k \in K, \tag{3}$$

$$\sum_{j_1 \in J} \sum_{j_2 \in J} (1 - \lambda_{kj_1}) x^k_{j_1 j_2} = 0, \ \forall k \in K, \tag{4}$$

$$\sum_{k \in K} \sum_{j_1 \in J} x^k_{j_1 j_2} \leq p_j, \ \forall j_2 \in J, \tag{5}$$

$$\sum_{z \in Z} \sum_{i \in I} (1 - q) q^{z-1} d_i w^{zk}_i \leq W, \ \forall k \in K, \tag{6}$$

$$\sum_{z \in Z} w^{zk}_i \leq 1, \ \forall k \in K, \ \forall i \in I, \tag{7}$$

$$\sum_{k \in K} w^{zk}_i = 1, \ \forall z \in Z, \ \forall i \in I, \tag{8}$$

$$y^{zk}_{ij_2} \leq \sum_{j_1 \in J} x^k_{j_1 j_2}, \ \forall z \in Z, \ \forall i \in I, \ \forall k \in K, \ \forall j_2 \in J, \tag{9}$$

$$w^{zk}_i = \sum_{j \in J} y^{zk}_{ij}, \forall i \in I, \forall z \in Z, \forall k \in K. \tag{10}$$

$$x^k_{j_1 j_2} \in \{0, 1\}, \ \forall j_1 \in J, \ \forall j_2 \in J, \ \forall k \in K, \tag{11}$$

$$w^{zk}_i \in \{0, 1\}, \ \forall z \in Z, \ \forall i \in I, \ \forall k \in K, \tag{12}$$

$$y^{zk}_{ij} \in \{0, 1\}, \ \forall z \in Z, \ \forall i \in I, \ \forall k \in K, \ \forall j \in J. \tag{13}$$

The ARDP seeks to minimize the total expected response time, as well as the time needed to perform the relocation (17). Constraints (3) and (4) aim to ensure that each ambulance is located somewhere, and constraint (5) guarantees that the limited number of ambulances that can be located at a standby site is satisfied. The maximal workload of ambulances is ensured through constraints (6): the expected number of demands assigned to an ambulance must be smaller or equal to its capacity. The expected workload assigned to an ambulance will depend on the busy fraction and the demand zones for which it is the first, the second, the third, to respond to calls. The model also takes into account that an ambulance cannot occupy more than one position on a given pre-assignment list (7), and that exactly one ambulance is assigned to each position on a given pre-assignment list (8). Constraints (9) and (10) ensure the link between decision variables while constraints (11) to (13) force their integrality.

## 4. Matheuristic decomposition approach

During a preliminary model analysis, the ARDP was solved by the means of a commercial solver (CPLEX 12.5). This tool was able to provide good quality solutions for small and medium size instances, including up to 150 demand zones and 12 ambulances. However, real-life instances become quickly challenging to solve as their size increases. We therefore developed a matheuristic decomposition approach to deal with larger size instances such as the ones considered in this study that includes up to 595 demand zones and 60 ambulances. The proposed solution approach is described in the current section.

Matheuristics consist of heuristic algorithms that include interoperation of metaheuristics and mathematic programming techniques. Indeed, they exploit features and solutions derived from the mathematical model of the problems of interest in some parts of the algorithms; hence, they are also called *model-based metaheuristics* (Boschetti et al., 2009; Maniezzo et al., 2009). In the literature, they have been applied to several problems, which are also close to the one addressed in this paper, e.g. the routing problem, as documented in a recent survey Archetti and Speranza (2014). According to Archetti and Speranza (2014), matheuristics can be classified into three classes: decomposition approaches, improvement heuristics, and column generation-based approaches. As we will see later on, our approach lies somewhere between the first and the second class. It is a decomposition approach because the problem is divided into smaller and simpler subproblems, and a specific model is applied to each subproblem. It is also an improvement heuristic since mathematical programming models are used to improve a solution found by a different heuristic approach; however, in our case, also this previous solution comes from a programming model. In the healthcare field, matheuristics are not largely exploited, yet. A research on scopus in July 2015 for journal papers including both *matheuristics* and *healthcare*

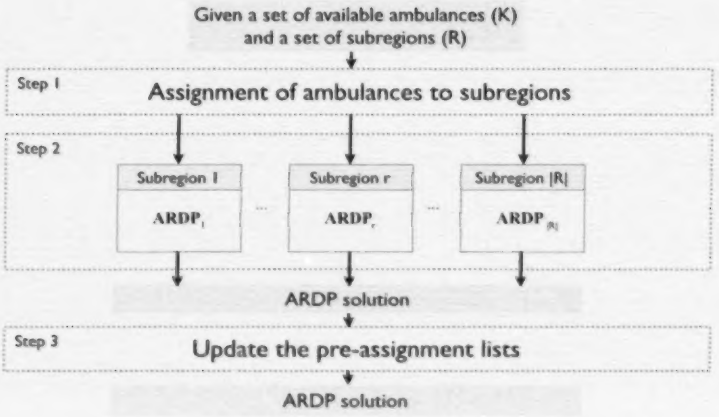


Figure 1: Architecture of the matheuristic decomposition approach

(or, alternatively, *ambulances*, only *care*, *doctor*, *emergency*, only *health*, *hospital*, *nurse*, *patient* or *physician*) as keywords only provided two results: Della Croce and Salassa (2014) and Caserta and Vobeta (2014). Nevertheless, only Della Croce and Salassa (2014) deals with a management problem (the nurse rostering problem), whereas Caserta and Vobeta (2014) develops an algorithm for the DNA sequencing problem.

The matheuristic decomposition approach developed to solve the ARDP exploits the division of the territory into subregions. The main idea of this approach comes from the fact that it is very unlikely to move an ambulance from one part of a city to a very far one, or else to dispatch an ambulance to serve a call that is very distant from its current location. Consequently, considering each subregion separately should provide good quality solutions. It also seems to be in line with several EMS organizations that choose to decompose the territory to serve into subregions or districts, and then manage them independently or with some sort of collaboration among the subregions. However, this also raises difficult questions regarding the definition of subregions as well as the collaboration process among the subregions.

Without loss of generality, in the context under study, we assume that decisions are taken in a centralized manner so any ambulance can serve any demand. The division of the territory into subregions is used for computational purpose rather than for management one. In our case, we assume that subregions are known a priori and based on the knowledge of the territory. Surely, subregions can be determined in several ways, but the best way of defining districts is a challenging problem that goes beyond the scope of this paper. As depicted in Figure 1, given a pre-determined set of subregions, the matheuristic decomposition approach proposed to solve the ARDP consists of three main steps: (1) ambulances are allocated to subregions; (2) the ARDP is solved for each subregion; (3) the pre-assignment lists are updated based on ambulance locations obtained from (2). Each step consists of a programming model extracted from the model presented in the previous section.

### *4.1. Step 1: Allocation of ambulances to sub-regions*

The first step of the solution approach aims to determine ambulance locations at an aggregated level, i.e in which subregions each ambulance should be located. It also intends to assign emergency calls to available ambulances. The objective of Step 1 is to define the best possible relocation plan such that both the total time to respond to calls and the total time to perform relocations are minimized. This objective function can then be formulated as:

$$\min \xi_s \sum_{i\in I}\sum_{z\in Z}\sum_{k\in K}\sum_{j\in J} d_i t_{ji} y_{ij}^{zk} + \xi_r \sum_{k\in K}\sum_{j_1\in J}\sum_{j_2\in J} \theta^k t_{j_1 j_2} x_{j_1 j_2}^k. \tag{14}$$

The problem solved in Step 1 is in fact an aggregated and simplified version of the ARDP, where (14) is taken as the objective function rather than (17). In this aggregated problem, we assume that an ambulance is always available to answer to a call: only one ambulance will be included in pre-assignment lists. Indeed, since several

ambulances are generally located at the same subregion (if the demand justifies it), the overall capacity of the ambulances allocated to a specific subregion should be able to serve the demand: idle ambulances will compensate for the non-availability of others. Consequently, constraints (6) to (8) are replaced by the following ones to adequately model to problem solved during Step 1:

$$\sum_{z\in Z}\sum_{i\in I} d_i w_i^{zk} \leq W,\ \forall k \in K, \tag{15}$$

$$\sum_{k\in K}\sum_{z\in Z} w_i^{zk} \leq 1,\ \forall i \in I. \tag{16}$$

A more refine assignment of demand zones to ambulances in a particular subregion is rather performed during Step 2. It is worth noting that assignment variables ($w_i^{zk}$) are now defined as the proportion of demands from $i$ that is assigned to ambulance $k$ at position $z$ on the pre-assignment list of $i$, and linking constraints as ($y_{ij}^{zk}$) as the proportion of demands from $i$ assigned to ambulance $k$ located at $j$ at position $z$ on the pre-assignment list of $i$. Both the assignment and linking variables are thus continuous between 0 and 1 rather than binary. As shown on Figure 1, solving the model of Step 1 will allow to define the set of ambulances, noted $K_r$, that will be located at each subregion $r \in R$, where $R$ is the set of considered subregions. The set $K_r$ will then be used to formulate each subproblem of Step 2.

### *4.2. Step 2: Ambulance relocation and pre-assignment for each subregion*

The second step is the core of the methodology. It corresponds, in fact, to solving several versions of the ARDP, one for each subregion $r \in R$. The solution to each subproblem provides the location of ambulances within a particular subregion as well as the pre-assignment list for each demand zone included in that subregion, both identified as $(x, w)_r$ on Figure 1. The combination of solutions found for each subregion $r \in R$ will provide an upper bound to the problem under study. When only one subregion is considered, i.e. $|R| = 1$, Step 2 is equivalent to the original problem. At this stage, each subproblem is independent. Consequently, to reduce the overall computation time, all subproblems of Step 2, noted ARDP$_r$, can be solved in parallel.

### *4.3. Step 3: Update of pre-assignment lists*

Any such decomposition approach can lead to significant problems at subregion boundaries. To overcome this issue, we propose to trigger a third phase in order to update pre-assignment lists considering fixed ambulance locations, i.e. the ones selected after Step 2. Doing so, it will be possible to find optimal pre-assignment lists given a set of pre-determined locations. To this end, all demand zones are taken into account simultaneously while solving the problem thus improving the quality of the upper bound found after Step 2. The problem solved during Step 3 is very similar to the original one where relocation decisions are assumed to be known. However, the objective function is slightly different: no relocation costs are involved in this step. The objective function of Step 3 is then:

$$\min \sum_{i\in I}\sum_{z\in Z}\sum_{k\in K}\sum_{j\in J}(1-q)q^{z-1}d_i t_{ji} y_{ij}^{zk}. \tag{17}$$

Moreover, since ambulance locations are known, only constraints (6) to (8) and (12) are included in the model.

## 5. Computational experiments

Computational experiments have been conducted to analyze both the ARDP and the matheuristic decomposition approach proposed to solve it. These experiments allow firstly to discuss the impact of integrating dispatching and relocation decisions in a common framework, as well as the impact of considering workload constraints and relocation costs on the decision-making. Secondly, they allow to evaluate how the decomposition of the problem into subproblems affects the final solution. The two sets of instances used throughout computational experiments are described hereafter.

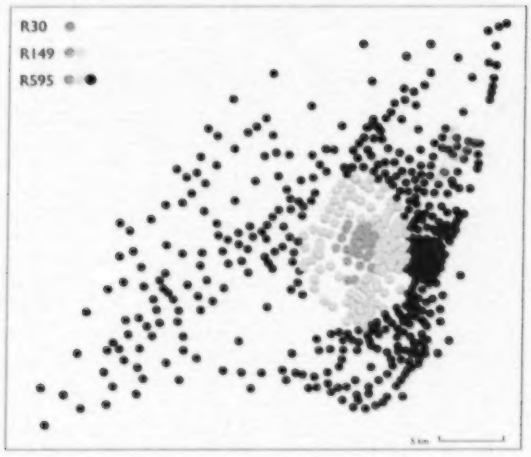


Figure 2: Schematic representation of instances

### *5.1. Instances description*

The first group of instances (referred to as random instances) has been generated randomly on a 20 km x 16 km rectangle territory. Random instances include 24 demand zones to each of which corresponds a demand $d_i$ varying with respect to the two following cases. In the first case, noted $U$, $d_i$ follows a uniform distribution between 1 and 100. The goal here is to replicate a urban-related case where lower and highly populated area exists all together. In the second case, noted $R$, $d_i$ is uniform and set to 50 for all demand zones. This set of instances aims to represent a rural context where the demand is more uniformly spread over the territory to serve. In both cases, we assume that each demand zone includes one standby site where an ambulance can be located. The travel time between each demand zone, or alternately, standby sites, are computed using the Euclidian distance and a constant speed of 50 km/h. To perform a broader model analysis, we consider cases where 3 and 4 ambulances are available to serve calls. In the base case, the maximal workload of ambulance is set to 425 interventions. This value has been set according to the generated demand profile, i.e. $d_i$. Moreover, the busy fraction has been set to 0.4. Other values have also been considered in the analysis and will be discussed when necessary.

The second group of instances (referred to as real-life inspired instances) has been created based on a real-life case. This set of instances is divided into three groups with respect to the size of instances: R30 (30 demand zones), R149 (149 demand zones) and R595 (595 demand zones). Regardless of the problem size, all real-life instances have been created based on the context of Montreal and Laval (the suburb just north of Montreal), which constitutes the major population centre in the province of Québec (Canada) with about 2.3 million inhabitants on a 744 square-km territory. This group of instances will allow to validate the methodology on a more realistic and practical setting. Figure 2 illustrates the whole region considered in this case (R595), as well as instance R30 and R149.

In all aforementioned cases, the expected demand (in terms of the expected number of interventions per time period) is extracted from simulated data (see Kergosien et al. (2015)). The number of potential standby sites with one ambulance capacity was arbitrary set from 30 to 218 with respect to the size of the considered instance. The travel time between each demand zone or standby site is computed using real travel time between centroids of zones. The number of available ambulances varies from 2 to 60 corresponding to cases where the number of zones per ambulance ranges from 7.5 to 20. Finally, for the base case, the maximal workload of ambulance is set to 5 interventions per 6-hour period whereas the busy fraction is set to 0.5.

The proposed matheuristic decomposition approach has been implemented within a Visual Basic framework, which includes a graphical interface that allows the user to easily modify the data and problem parameters. The VB software also manages all the data needed at each step of the matheuristic and interacts with OPL 5.1 to build the required models. Then, each subproblem within the matheuristic decomposition approach is solved by the means of CPLEX 12.5. Each computational test was run on a Microsoft Windows

| *Inst.* | $\|R\|$ | $\|I\|$ | $\|J\|$ | $\|K\|$ | $\|Z\|$ | $d_k$ | q | W | $\xi_a$ | Section |
|---|---|---|---|---|---|---|---|---|---|---|
| U | 1 | 24 | 24 | 3,4 | 1,2,3,4 | UNIF[1,100] | 0.4 | 425 | 1 | *5.1.1* |
| U | 1 | 24 | 24 | 3,4 | 2 | UNIF[1,100] | 0.4 | 400, 425, 500, 2400 | 1 | *5.1.2* |
| R | 1 | 24 | 24 | 3,4 | 1,2,3,4 | 50 | 0.4 | 425 | 1 | *5.1.1* |
| R | 1 | 24 | 24 | 3,4 | 2 | 50 | 0.4 | 400, 425, 500, 2400 | 1 | *5.1.2* |
| R30 | 1 | 30 | 30 | 2,3,4 | 1,2,3,4 | Simul. | 0.5 | 5 | 1 | *5.1.1* |
| R30 | 1 | 30 | 30 | 2,3,4 | 2 | Simul. | 0.5 | 2, 3, 4, 5, 2400 | 1 | *5.1.2* |
| R30 | 1 | 30 | 30 | 2,3,4 | 2 | Simul. | 0.5 | 5 | 1,0.875,0.75,0.5,0 | *5.1.3* |
| R30 | 1,2 | 30 | 30 | 4 | 2 | Simul. | 0.5 | 5 | 1 | *5.2* |
| R149 | 1 | 149 | 48 | 10 | 1,2 | Simul. | 0.5 | 5 | 1 | *5.1.1* |
| R149 | 1 | 149 | 48 | 8,10,12,14 | 1,2 | Simul. | 0.5 | 5 | 1 | *5.1.2* |
| R149 | 1 | 149 | 48 | 10 | 2 | Simul. | 0.5 | 5 | 1,0.75,0.5,0.25,0 | *5.1.3* |
| R149 | 1,2,3,4 | 149 | 48 | 10 | 2 | Simul. | 0.5 | 5 | 1 | *5.2* |
| R595 | 15 | 595 | 218 | 50 | 1,2 | Simul. | 0.5 | 5 | 1 | *5.1.1* |
| R595 | 15 | 595 | 218 | 40, 50, 60 | 2 | Simul. | 0.5 | 5 | 1 | *5.1.2* |
| R595 | 15 | 595 | 218 | 50 | 2 | Simul. | 0.5 | 5 | 1,0.75,0.5,0.25,0 | *5.1.3* |
| R595 | 5, 10, 15 | 595 | 218 | 50 | 2 | Simul. | 0.5 | 5 | 1 | *5.2* |

Table 2: Summary of the main instances

machine with eight cores and 15 GB RAM, which was installed on a server equipped with an AMD Opteron 6328 processor. Overall, over 250 different cases were solve to get a detailed analysis of both the model and the solution approach. However, we will only report and discuss in the following sections the most relevant results to the purpose of this paper. A summary of the main instances are summarized in Table 2 along with their corresponding sections. Note that all tested instances are not presented in the table, but only the ones that are discussed in the paper.

### *5.2. Model analysis*

Hereafter, we focus on three aspects of the problem that we deem to be the most important to discuss to draw a good appraisal of the ARDP itself. The discussion includes: (1) the impact of pre-assignment lists, (2) the impact of workload capacity, and (3) the analysis of relocation costs. Each of the main results is presented in a table form where the main characteristics of the instance solved are reported, together with the final solution both in terms of location decisions (LOC), the total expected response time ($ERT^T$) and the expected response time per intervention (ERT), the computational time (CT), and the gap between the best solution and the lower bound found using CPLEX (GAP). When feasible, we consider only one subregion, i.e. the original problem is solved, with a imposed time limit of 28 800 seconds; larger cases are faced with the matheuristic decomposition approach with $|R| > 1$.

#### *5.2.1. Pre-assignment lists*

Table 3 reports results obtained for smaller size instances, namely *U*, *R* and *R30*, for various values of $|Z|$, the list size. All reported cases consider that four available ambulances are available to answer to calls. However, a similar behaviour has been observed for other fleet size. To better assess the impact of pre-assignment lists, in the following experiments, we did not consider any relocation cost, i.e. $\xi_r = 0$. Relocation costs will rather be analyzed in section 5.2.3.

| | $\|Z\|$ | LOC | $ERT^T$ [1000 s] | ERT [s/int.] | CT [s] | GAP [%] |
|---|---|---|---|---|---|---|
| *U* | 1 | [5, 8, 16, 19] | 575.6 | 428 | 143 | 0.00 |
| | 2 | [8, 11, 14, 17] | 532.0 | 399 | 276 | 0.00 |
| | 3 | [8, 11, 14, 16] | 530.5 | 397 | 28 800 | 0.00 |
| | 4 | [8, 11, 14, 16] | 530.5 | 397 | 28 800 | 0.00 |
| *R* | 1 | [2, 5, 14, 17] | 565.9 | 423 | 2 | 0.00 |
| | 2 | [8, 11, 14, 17] | 490.8 | 367 | 793 | 0.00 |
| | 3 | [8, 11, 14, 17] | 490.8 | 367 | 28 800 | 12.99 |
| | 4 | [8, 11, 14, 17] | 501.2 | 375 | 28 800 | 18.27 |
| *R30* | 1 | [2, 6, 20, 30] | 1.89 | 281 | 2 | 0.00 |
| | 2 | [2, 11, 14, 20] | 1.82 | 271 | 1731 | 0.00 |
| | 3 | [2, 7, 11, 20] | 1.82 | 271 | 28 800 | 8.76 |
| | 4 | [2, 7, 11, 20] | 1.82 | 271 | 28 800 | 9.70 |

Table 3: Impact of pre-assignment lists

Three main observations arise from results presented in Table 3. First of all, in all studied cases, pre-assignment lists have a significant impact on location decisions, particularly when two ambulances are included

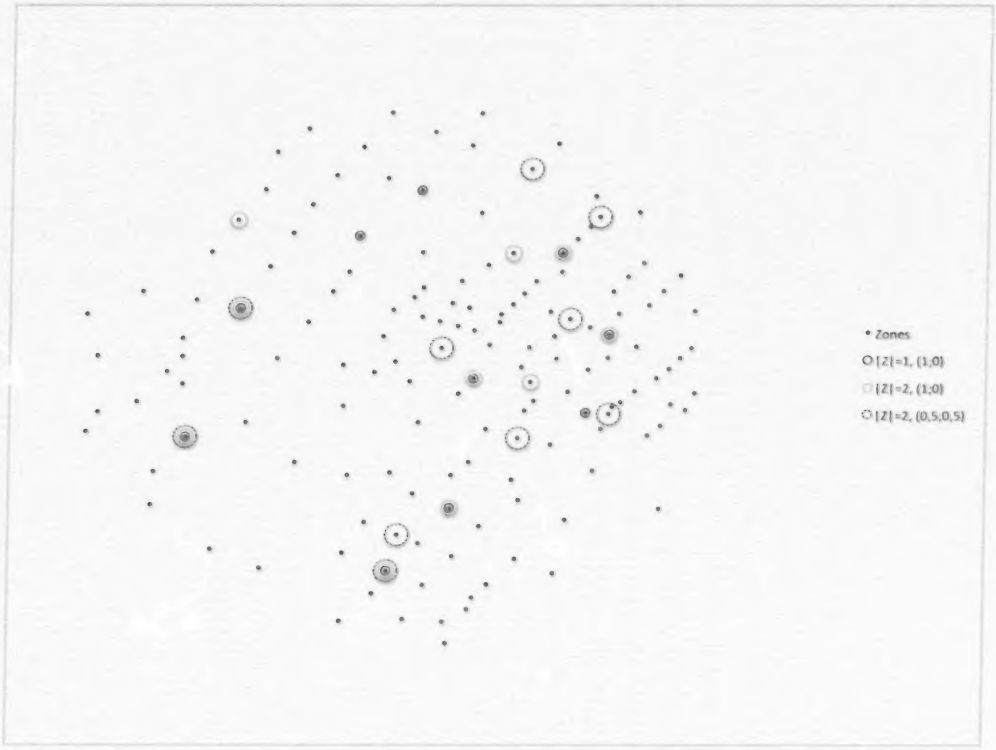


Figure 3: Results - R149 - 10 ambulances

in pre-assignment lists rather than one. Since the marginal contribution of each supplementary ambulance reduces in the objective function with the value of $|Z|$, smaller variations in terms of location decisions are obtained afterwards. Secondly, the expected performance of the system are improving when pre-assignment lists are taken into account. Given the instance, the improvement can go up to 56 seconds per intervention when two ambulances are included on pre-assignement lists instead of one. The marginal improvement then reduces as the value of $|Z|$ increases. Consequently, we believe that, in our cases, considering two ambulances on the list seems to be sufficient enough. This observation is also in line with the widely spread idea of double coverage (Hogan and ReVelle, 1986; Gendreau et al., 1997). Nevertheless, the expected response time measure allows a more refine differentiation among equivalent solutions from the coverage perspective. Finally, computational time increase significantly with the size of the list. For these reasons, we will set the value of $|Z|$ to 2 which constitutes, in our opinion, a fair tradeoff between service level improvement and computational time.

Figure 3 shows solutions found for R149 with several parameter settings. In those cases, although optimal solutions cannot be found within the time limit, the general behaviour observed for smaller instances still hold: pre-assignment lists impact location decisions. In addition, despite the challenge in finding solutions to the original version of R595, results obtained by the means of the proposed matheuristic decomposition approach for $|R| > 1$ also seem to confirm this observation.

*5.2.2. System capacity*

The overall capacity of the system varies according to two main parameters: the number of ambulances $|K|$ and their maximal workload $W$. Different combinations of $|K|$ and $W$ allow us to define several scenarios with respect to the system capacity, and consequently, to analyze how the system capacity will influence both location and dispatching decisions. Table 4 reports the solution for instances $U$, and for varying system capacity. However, similar behavior are observed for all instances. In this case, $W = 2400$ corresponds to a

unlimited system capacity. In addition to those presented previously, we also provide additional performance measures. The standard deviation of the expected ambulance workload with and without the list (where ambulances are always ordered according to an increase distance criteria), noted $\sigma^Z$ and $\sigma^N$, respectively, will allow to better evaluate the workload balance for each scenario whereas the percentage of demands that are not covered by the nearest ambulance, noted $\%_D$, will highlight the potentiel loss in the expected response time.

| | $\|K\|$ | $W$ | LOC | $\mathbf{ERT}^T$ [1000 s] | ERT [s/int] | $\sigma^Z$ | $\sigma^N$ | $\%_D$ |
|---|---|---|---|---|---|---|---|---|
| | 3 | 400 | [11, 14, 16] | 599.8 | 449 | 43.5 | 83.8 | 7.0 |
| | 3 | 425 | [11, 14, 16] | 597.3 | 447 | 80.3 | 83.8 | 7.0 |
| | 3 | 500 | [11, 14, 16] | 596.9 | 446 | 83.8 | 83.8 | 0.0 |
| $U$ | 3 | 2400 | [11, 14, 16] | 596.9 | 446 | 83.8 | 83.8 | 0.0 |
| | 4 | 400 | [8, 11, 14, 17] | 532.9 | 399 | 27.7 | 27.7 | 0.0 |
| | 4 | 425 | [8, 11, 14, 17] | 532.9 | 399 | 27.7 | 27.7 | 0.0 |
| | 4 | 500 | [8, 11, 14, 17] | 532.9 | 399 | 27.7 | 27.7 | 0.0 |
| | 4 | 2400 | [8, 11, 14, 17] | 532.9 | 399 | 27.7 | 27.7 | 0.0 |

Table 4: Impact of system's capacity

The main results obtained show that the ambulance capacity does not have a clear impact on location decisions. When we compare results obtained for scenarios in which the same value of $|K|$ is considered, but where different values of $W$ are taken into account, location decisions remain the same. However, when the maximal ambulance workload decreased, pre-assignment lists are modified to account for the reduced ambulance capacity. This allows a better estimate of system performances as well as ensuring a better workload balance. Indeed, the variability in this expected workload assigned to each ambulance seems to decrease with the ambulance capacity. The variability decreases by up to 48 % for an increase of less than 0.5% in the expected response time. Not considering workload constraints can thus have a non-negligeable impact on workload balance. Finally, unsurprisingly, in all cases, system performances are improving with the number of ambulances.

*5.2.3. Relocation costs*

All previous tests were run considering only the minimization of the total expected response time: no relocation costs were included in the objective function, i.e. $\xi_r = 0$. In this section, we include relocation costs in the objective function allowing the analysis of the tradeoff between service improvement and relocation efforts. Several scenarios have been created with respect to $\xi_s$ and $\xi_r$: values of $\xi_s$ have been set to 1, 0.875, 0.75, 0.5 and 0, with $\xi_s + \xi_r = 1$. In what follows, we will present detailed results obtained for R30, for several ambulance location plans before the relocation is triggered. Indeed, ambulance locations prior to the relocation will significantly influence relocation decisions. To represent varied situations, ambulance locations before the relocation were determined in three ways: locations are selected randomly, or locations chosen are in the subset of the optimal locations when $|K| + 1$ ambulances are available to answer to calls, or a combination of both. Figure 4 reports results obtained for 2, 3 and 4 ambulances, respectively. For each figure, the total expected response time is plotted as a function of the total relocation time. This has been done for several ambulance location plans before the relocation is triggered (location plans before relocation are identified between brackets, each number corresponds to the location of an ambulance). Each point on the graph represents the solution found for a given combination of $\xi_s$ and $\xi_r$. It is important to note that, again, although we limit our discussion to the results obtained for R30 for the various cases described here above, the general behaviour of the system in this particular case reflects what we obtained for other instances and parameter settings.

First of all, results obtained for several values of $\xi_s$ and $\xi_r$ allow to conclude that, unsurprisingly, the weights given to both objective impact significantly location decisions. It is also possible to see that, when the value of $\xi_s$ is reducing, quickly, performance improvement are not sufficient to justify relocations efforts. Secondly, the location of ambulances prior to the relocation also notably impacts on relocation decisions. Performance improvement are generally less justified when the initial location plan is close to the optimal plan for $|K| + 1$, meaning that, for instance, a relocation process has just been launched. Indeed, when all initial standby sites belong to the set of optimal standby points for $|K| + 1$, the current solution, without any relocation, is still of a good quality. In these cases, we can observe that the difference between the current and the optimal solutions ranges from 1.4% to 8.7%, depending on the number of available ambulances. For

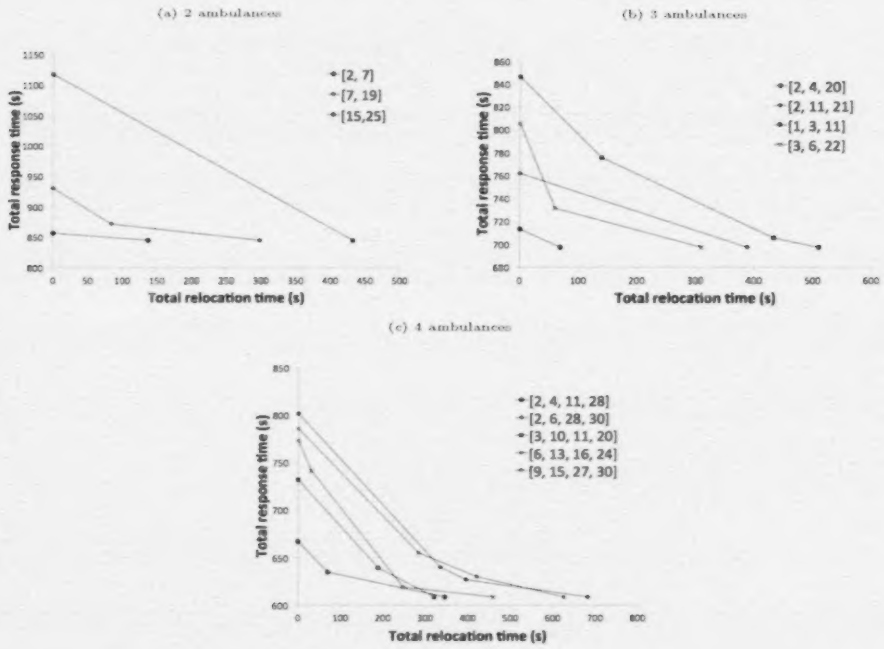


Figure 4: Total expected response time vs total relocation time

a practical perspective, it becomes really interesting to modify the current standby sites of available vehicles when the current location plan significantly differs from the optimal one for $|K| + 1$, for instance, when the last relocation process occurs a while ago. Moreover, when the number of ambulances is larger, and yet the system has more flexibility, the *statu quo* strategy become less and less interesting, at least from a service level standpoint. Finally, a thorough look at Figure 4 allow us to conclude that the marginal improvement is considerable when slight modifications to the current location plan are accepted. Then, the marginal gain tends to decrease as the current solution is modified. It is thus interesting to relocate available ambulances to improve the service level, particularly in the aforementioned case, but the additional efforts in terms of relocation times are not always justify after a certain point. Figure 3 shows that, for R149, different locations plan are obtained when $\xi_s = 0.5$. This is also the case for R595 when solved by the means of the matheuristic decomposition approach with $|R| = 15$.

### *5.3. Matheuristic analysis : the impact of decomposing the problem*

To validate the matheuristic decomposition approach, real-life inspired instances have been decomposed into several subregions (see Figure 5). The number of subregions, $|R|$, and the decomposition patterns have been determined based on the geographical characteristics of the territory and such that a fairly uniform territory division can be obtained, in terms of number of zones per subregions. For R149 and R595, we divided the territory into subregions so that ratios ranging from 37.5 to 120 zones per subregion can be obtained. This ratio is significantly smaller for R30, namely 15 zones per subregion, due to its limited size. However, in this case, two decomposition patterns were tested. In the following sections, we will address the decomposition itself, i.e. how much we loose in terms of the quality of the final solution, and discuss the computational time and the contribution of Step 3 to the quality of the final solution. Again, only the main results are presented here and reflect what we obtained for other instances.

Table 5 presents results obtained for several decomposition patterns. For each group of instances and value of $|R|$, we report both the value of the objective function (OF) found after Step 2 and Step 3. The computational time, noted CT, and the average GAP for each of the subproblem solved during Step 2, noted $\overline{GAP}_{S2}$, are also provided in this table. For each case, a time limit of 3600 seconds was imposed and pre-assignment lists including 2 ambulances have been taken into account. The best solution found for each group of instances are highlighted in the table.

(a) $|R| = 5$

(b) $|R| = 10$

(c) $|R| = 15$

Figure 5: Decomposition patterns - R595

| | $\|R\|$ | Step | OF [s] | CT [s] | $\overline{GAP}_{S2}$ [%] |
|---|---|---|---|---|---|
| *R30* | 1 | 1+2 | **608.6** | 1731 | 0 |
| | 1 | 1+2+3 | **608.6** | 1731 | 0 |
| | 2-I | 1+2 | 778.4 | 2 | 0 |
| | 2-I | 1+2+3 | 659.7 | 2 | 0 |
| | 2-II | 1+2 | 631.37 | 2 | 0 |
| | 2-II | 1+2+3 | **608.6** | 2 | 0 |
| *R149* | 1 | 1+2 | 5886.9 | 3600 | 24.6 |
| | 1 | 1+2+3 | 5886.9 | 3600 | 24.6 |
| | 2 | 1+2 | 5563.3 | 3600 | 9.82 |
| | 2 | 1+2+3 | **5320.3** | 3600 | 9.82 |
| | 3 | 1+2 | 5724.9 | 2610 | 2.96 |
| | 3 | 1+2+3 | 5460.5 | 2610 | 2.96 |
| | 4 | 1+2 | 5871.4 | 203 | 0 |
| | 4 | 1+2+3 | 5461.0 | 203 | 0 |
| *R595* | 1 | 1+2 | - | - | - |
| | 1 | 1+2+3 | - | - | - |
| | 5 | 1+2 | 37569.1 | 3600 | 33.12 |
| | 5 | 1+2+3 | 28381.3 | 3600 | 33.12 |
| | 10 | 1+2 | 31165.1 | 3561 | 11.22 |
| | 10 | 1+2+3 | 28097.0 | 3561 | 11.22 |
| | 15 | 1+2 | 31161.7 | 1174 | 1.01 |
| | 15 | 1+2+3 | **28929.0** | 1174 | 1.01 |

Table 5: Analysis of the matheuristic decomposition approach

First of all, results show that the decomposition pattern affects the final solution. For R30, the optimal solution can be found using the second decomposition pattern whereas a fair solution (within 10% of the optimal one) can be obtained in the first case. For R149, given the time limit, the matheuristic decomposition approach is always able to find a better solution than the one obtained for the global problem, i.e. when $|R| = 1$. For real-life instances, the matheuristic decomposition approach is able to provide a solution when $|R| > 1$, which is not the case when $|R| = 1$. Good quality solutions can be determined using the decomposition approach. Unsurprisingly, computational times are considerably reduced when the number of subregions increases.

Results provide in Table 5 allow us to conclude that the decomposition is an effective approach, both in terms of solution quality and computational time. However, there is also a tradeoff between the quality of the solution obtained for each subproblem and the quality of the global solution. We observe that, for R149, smaller is the number of subregions, better is the solution. Indeed, when the number of subregions is smaller we are closer to the global problem. However, in some cases, like R595, resulting subproblems remain very large when the value of $|R|$ is smaller. Consequently, we observe the opposite behaviour for the larger size instances: larger is the number of subregions, better is the solution.

Finally, it is also worth noting that, in all studied cases, Step 3 significantly helps improve the quality of the solution. We also observe that, thanks to Step 3, reducing the time limit up to 300 seconds lead to better or similar results that the one found imposing a time limit of 3600 seconds. In fact, the best solution after Step 3 was found imposing a time limit of 900 seconds, although the solution found after Step 2 continues to improve slightly with the time limit.

## 6. Conclusion

This article focuses on the modelling and the analysis of a decision model for EMS management that jointly addresses relocation and dispatching decisions. The ARDP as proposed in this study differs from other related problems by several aspects. First, in addition to relocation decisions, it incorporates the establishment of ordered pre-assignment lists that guide dispatching decisions. It also considers explicitly the expected ambulance workload and uses the expected response time to select the best possible solution and assess system performances.

Results show that including more than one ambulances in the pre-assignment lists influences location decisions and, consequently, the system performances. The capacity of ambulance, i.e. their maximal workload, also impact the decisions, but mostly pre-assignment ones. Indeed, when the system capacity is tight, pre-assignment lists adjust accordingly thus providing a better estimate of system performances and workload balance. Finally, the weights assigned to objectives related to both service to the population and relocation efforts also influence decisions. These findings have been validated for smaller size instances, but also for larger ones solved by the means of the proposed decomposition matheuristic. The matheuristic approach has

indeed proven to be an effective tool to solve real-life instances in a timely manner. The solution approach was able to find the optimal solution for smaller instances and improves significantly the solution determined for medium size instances using comparable time limit. This paper therefore provide both a model and a methodology to deal with relocation and pre-assignment decisions in real-life settings, and opens up promising research perspectives.

The proposed solution approach has been implemented in a decision-support tool that can help deal with such decisions but can also support the analysis of other types of decisions such as fleet dimensioning and districting. As it is right now, the matheuristic decomposition approach has been implemented such that the steps are solved sequentially. Although very good results were obtained through the use of this methodology, in our opinion, it is likely that it can greatly benefit from the introduction of different feedback mechanisms and local search inspired techniques. The refinement of the proposed methodology constitutes one of the research avenues that we want to address. Moreover, we believe that the idea of pre-assignment lists raised in this paper could be transposed in other contexts where the availability of servers is uncertain.

## References


Andersson, T., Värbrand, P., 2007. Decision support tools for ambulance dispatch and relocation. Journal of the Operational Research Society 58, 195–201.

Archetti, C., Speranza, M. G., 2014. A survey on matheuristics for routing problems. EURO Journal on Computational Optimization 2, 223–246.

Başar, A., Çatay, B., Ünlüyurt, T., 2011. A multi-period double coverage approach for locating the emergency medical service stations in Istanbul. Journal of the Operational Research Society 62, 627–637.

Başar, A., Çatay, B., Ünlüyurt, T., 2012. A taxonomy for emergency service station location problem. Operations Research Letters 6, 1147–1160.

Bandara, D., Mayorga, M. E., McLay, L. A., 2012. Optimal dispatching strategies for emergency vehicles to increase patient survivability. International Journal of Operational Research 15, 195–214.

Bandara, D., Mayorga, M. E., McLay, L. A., 2014. Priority dispatching strategies for EMS systems. Journal of the Operational Research Society 65, 572–587.

Bélanger, V., Kergosien, Y., Ruiz, A., Soriano, P., 2014. An empirical comparison of relocation strategies in real-time ambulance fleet managament. Tech. Rep. CIRRELT-2014-73, CIRRELT.

Bélanger, V., Ruiz, A., Soriano, P., 2012. Déploiement et redéploiement des véhicules ambulanciers dans la gestion des services préhospitaliers d'urgence. INFOR 50, 1–30.

Bélanger, V., Ruiz, A., Soriano, P., 2015. Recent advances in emergency medical services management. Tech. Rep. CIRRELT-2015-28, CIRRELT.

Boschetti, M. A., Maniezzo, V., Roffilli, M., Röhler, A. B., 2009. Matheuristics: Optimization, simulation and control matheuristics: Optimization, simulation and control hybrid metaheuristics. Lecture Notes in Computer Science 5818, 171–177.

Brotcorne, L., Laporte, G., Semet, F., 2003. Ambulance location and relocation models. European Journal of Operational Research 147, 451–463.

Carter, G. M., Chaiken, J. M., Ignall, E., 1972. Response areas for two emergency units. Operations Research 20, 571–594.

Caserta, M., Vobeta, S., 2014. A hybrid algorithm for the DNA sequencing problem. Discrete Applied Mathematics 163, 87–99.

Daskin, M. S., 1983. A maximum expected location problem : Formulation, properties and heuristic solution. Transportation Science 17, 416–439.

Della Croce, F., Salassa, F., 2014. A variable neighborhood search based matheuristic for nurse rostering problems. Annals of Operations Research 218, 185–199.

Gendreau, M., Laporte, G., Semet, F., 1997. Solving an ambulance location model by tabu search. Location Science 5, 75–88.

Gendreau, M., Laporte, G., Semet, F., 2001. A dynamic model and parallel tabu search heuristic for real-time ambulance relocation. Parallel Computing 27, 1641–1653.

Gendreau, M., Laporte, G., Semet, F., 2006. The maximal expected relocation problem for emergency vehicles. Journal of the Operational Research Society 57, 22–28.

Goldberg, J., 2004. Operations research models for the deployment of emergency services vehicle. EMS Management Journal 1, 20–39.

Hogan, K., ReVelle, C. S., 1986. Concepts and application of backup coverage. Management Science 34, 1434–1444.

Ingolfsson, A., 2013. EMS planning and management. In: Zaric, G. S. (Ed.), Operations Research and Healthcare policy. Springer, New York, N.Y., pp. 105–128.

Jagtenberg, C. J., Bhulai, S., van der Mei, R. D., 2015. An efficient heuristic for real-time ambulance redeployment. Operations Research for Health Care 4, 27–35.

Kergosien, Y., Bélanger, V., Soriano, P., Gendreau, M., Ruiz, A., 2015. A generic and flexible simulation-based analysis tool for EMS management. International Joural of Production Research, In Press.

Larson, R. C., 1974. A hypercube queuing model for facility location and redistricting in urban emergency services. Computers & Operations Research 1, 67–85.

Maniezzo, V., Stetzle, T., Voss, S. (Eds.), 2009. Matheuristics: Hybridizing Metaheuristics and Mathematical Programming. Vol. Annals of Information Systems. Springer, Heidelberg.

Mason, A. J., 2013. Simulation and real-time optimised relocation for improving ambulance operations. In: Denton, B. (Ed.), Handbook of Healthcare Operations: Methods and Applications. Springer, New York, N.Y., pp. 289–317.

Maxwell, M. S., Restepo, M., Henderson, S. G., Topaloglu, H., 2009. Approximate dynamic programming for ambulance redeployment. INFORMS Journal on Computing 22, 266–281.

McLay, L. A., Mayorga, M. E., 2013. A model for optimally dispatching ambulances to emergency calls with classification errors in patient priorities. IIE Transactions 45, 1–24.

Nair, R., Miller-Hooks, E., 2009. Evaluation of relocation strategies for emergency medical service vehicles. Journal of the Transportation Research Board 2137, 63–73.

Naoum-Sawaya, J., Elhedhli, S., 2013. A stochastic optimization model for real-time ambulance redeployment. Computers & Operations Research 40, 1972–1978.

Rajagopalan, H. K., Saydam, C., Xiao, J., 2008. A multiperiod set covering location model for dynamic redeployment of ambulances. Computers & Operations Research 35, 814–826.

Repede, J. F., Bernardo, J. J., 1994. Developping and validating a decision support system for location emergency medical vehicles in Louisville, Kentucky. European Journal of Operational Research 75, 567–581.

Saydam, C., Rajagopalan, H. K., Sharer, E., Lawrimore-Belanger, K., 2013. The dynamic redeployment coverage location model. Health Systems 2, 103–119.

Schmid, V., 2012. Solving the dynamic ambulance relocation problem and dispatching problem using approximate dynamic programming. European Journal of Operational Research 219, 611–621.

Schmid, V., Doerner, K. F., 2010. Ambulance location and relocation problems with time-dependent travel times. European Journal of Operational Research 207, 1293–1303.

Sudtachat, K., Mayorga, M. E., McLay, L. A., 2016. A nested-compliance table policy for emergency medical service systems under relocation. Omega 58, 154–168.

Toro-Diaz, H., Mayorga, M. E., Chanta, S., McLay, L. A., 2013. Joint location and dispatching decisions for emergency medical services. Computers & Industrial Engineering 64, 917–928.

Toro-Diaz, H., Mayorga, M. E., McLay, L. A., Rajagopalan, H. K., Saydam, C., 2015. Reducing disparities in large-scale medical service systems. Journal of the Operational Research Society 66, 1169–1181.

van den Berg, P. L., Aardal, K., 2015. Time-dependent MEXCLP with start-up and relocation cost. European Journal of Operational Research 242, 383–389.